FinAl: Fine-grained Alignment for Detail-Preserving Medical Vision-Language Pretraining

Jongsu Youn[1] **Dongyoung Lee**[1] **JaeHyeong Bae**[1]
Sang-Il Choi[2*] **Sangtae Choi**[3*] **Dae Ung Jo**[4*] **Jongwon Choi**[1,5,6*]
[1]Dept. of Advanced Imaging, GSAIM, Chung-Ang Univ., Republic of Korea
[2]Dept. of Computer Eng., Dankook Univ., Republic of Korea
[3]Dept. of Internal Medicine, Chung-Ang Univ. Coll. of Medicine, Seoul, Republic of Korea
[4]Sch. of Electronics Eng., Kyungpook Natl. Univ., Republic of Korea
[5]GS. of AI, Chung-Ang Univ., Republic of Korea
[6]Dept. of Metaverse Convergence, Chung-Ang Univ., Republic of Korea
{whyjs,ldy,bjhfirst}@vilab.cau.ac.kr
choisi@dankook.ac.kr beconst@cau.ac.kr
daeung.jo@knu.ac.kr choijw@cau.ac.kr

Abstract

We present FinAl, a medical vision–language pretraining (VLP) framework for preserving fine-grained clinical information such as disease severity. Hard one-hot image–report supervision does not explicitly encode graded relationships among clinically related non-paired samples. FinAl converts within-batch similarity rankings of structured radiology reports into soft targets for global image–text alignment, assigning graded target probabilities to non-paired samples without requiring severity annotations during pretraining. We combine this objective with local alignment and adapt the global prediction temperature according to disagreement between image–image and report–report similarity rankings. In zero-shot severity grading on EdemaSeverity and three CAD-Chest subsets, FinAl achieves the lowest mean absolute error and highest quadratic weighted kappa among the compared models. FinAl also hows gains in severity-aware retrieval and frozen-encoder linear probing while remaining competitive on category-level retrieval.

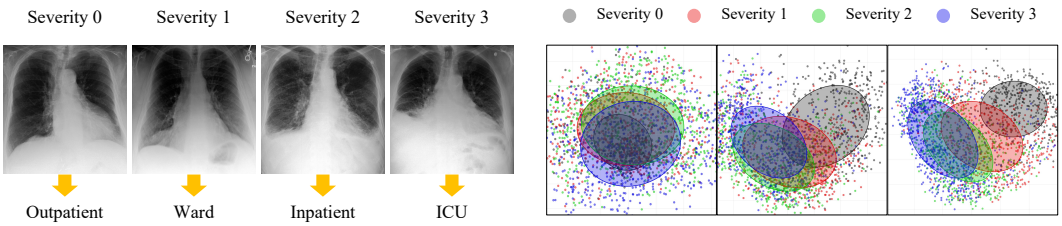
1 Introduction

Medical imaging models are often trained and evaluated on whether a disease is present, but many clinical applications also require estimating how severely a finding is expressed. Obtaining reliable ordinal severity labels requires consistent grading criteria and expert judgment to distinguish adjacent clinical states. This annotation burden motivates learning from the clinical information already available in radiology reports. Medical vision-language pretraining (VLP) uses paired chest X-ray images and reports to learn clinically useful multimodal representations without requiring severity annotations during pretraining, and has shown strong performance on medical classification and retrieval benchmarks [1, 2, 3]. However, conventional hard one-hot contrastive supervision assigns a target probability of one to the paired image–report match and zero to all non-paired candidates. While this objective learns image–text correspondence, its targets do not distinguish degrees of clinical relatedness among non-paired samples. For example, cases with adjacent disease severity levels or partially overlapping findings receive the same zero target probability as clinically distant cases. Consequently, the alignment objective does not explicitly encourage the learned representations to preserve these graded clinical relationships.

In clinical practice, patients with the same disease may require different treatment strategies and levels of care depending on disease severity. Figure 1(a) illustrates these severity distinctions using

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Chest X-rays at different pulmonary edema severity levels (b) SAT (c) MGCA (d) FinAl (ours)

Figure 1: Motivation for fine-detail preserving medical vision-language pretraining.

chest X-rays spanning an ordinal spectrum from no pulmonary edema to severe pulmonary edema. Capturing such gradations is therefore important for learning medical image representations that preserve clinically relevant differences beyond disease presence. On the CAD-Chest pulmonary edema benchmark, PCA visualizations of SAT and MGCA show substantial overlap among severity-wise distributions of visual embeddings (Fig. 1(b)–(c)). These qualitative observations motivate examining whether pretraining can better preserve severity-related structure. Radiology reports offer an indirect source of supervision for this goal because they describe clinical findings and their attributes. Our analysis on EdemaSeverity shows that mean report–report similarity decreases with increasing severity distance (Section 4.6), supporting the use of inter-report relationships as a signal of graded clinical relatedness.

To address this issue, we introduce FinAl, a medical vision–language pretraining framework that uses inter-report relationships to preserve fine-grained clinical information without severity annotations during pretraining. FinAl converts the within-batch ordering of report similarities into soft targets for global image–text alignment. These targets assign graded probabilities to non-paired samples according to their report-similarity ranks, rather than assigning zero target probability to every non-paired sample. By using ranks, target construction depends on the relative ordering of report similarities rather than their absolute magnitudes. We combine this global objective with local cross-attention alignment [4] for region–token correspondence. We also use modality-wise similarity discrepancy (MSD), computed from the disagreement between image–image and report–report similarity rankings within each batch, to adapt the temperature of the global image–text prediction distribution. Fig. 1(d) provides a qualitative illustration of the resulting severity-related structure in the visual embedding space.

Our goal is not to develop a task-specific severity classifier, but to improve the preservation of graded clinical differences in pretrained representations. Through zero-shot severity grading, severity-aware retrieval, and frozen-encoder linear probing on chest X-ray benchmarks, we show that FinAl learns representations that better distinguish graded clinical states. These findings support incorporating inter-report relationships into the alignment objective to preserve fine-grained clinical information in pretrained representations.

2 Related work

2.1 Vision-Language Pretraining for Medical Imaging

ConVIRT [1] established global contrastive learning from paired medical images and reports. GLoRIA [5] adds word–patch alignment, while MGCA [4] combines global, local, and disease-level objectives with bidirectional cross-attention. MedKLIP [2] grounds structured disease and anatomical entities, and CARZero [3] refines cross-modal similarity modeling for zero-shot inference. SAT [6] mitigates false negatives by thresholding inter-report similarities into positive, neutral, and negative groups. Recent methods further enrich medical VLP with structured clinical semantics. ASG [7] uses anatomical-region, finding, and existence triplets for clinically grounded local alignment. MAVL [8] decomposes disease descriptions into expert-curated visual aspects, whereas RADZero [9] extracts finding-level sentences and aligns them with local image features through similarity-based cross-attention. These methods primarily refine anatomical, aspect-level, or finding-level correspondence, while FinAl modifies the global alignment target to preserve graded inter-report relationships.

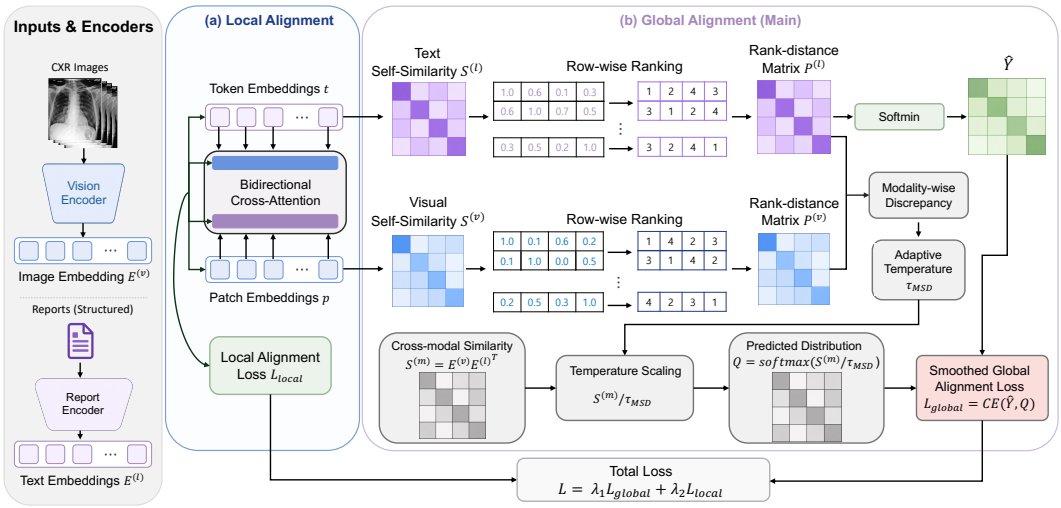


Figure 2: Overall framework of FinAl. (a) Local alignment encourages patch-token correspondence via bidirectional cross-attention. (b) Global alignment constructs rank-induced soft targets from report self-similarity and uses modality-wise rank discrepancy to adapt the temperature τ_{MSD}. Cross-modal similarities are computed directly from image and text embeddings, scaled by τ_{MSD}, and supervised with the smoothed global alignment loss.

FinAl instead preserves graded clinical differences by converting the full within-batch ordering of report similarities into soft global targets. Unlike SAT's fixed triage, FinAl assigns graded target probabilities to non-paired samples; Section 4.6 and Appendix D.10 analyze this distinction.

2.2 Soft Supervision and Rank-aware Representation Learning

Prior work addresses semantic overlap through different mechanisms. Debiased contrastive learning [10] corrects negative-sampling bias caused by same-class samples, while Probabilistic Cross-Modal Embedding (PCME) [11] models uncertainty and one-to-many correspondences through probabilistic embeddings. FinAl focuses on preserving graded clinical differences by encoding inter-report relationships directly in cross-modal alignment targets, rather than correcting negative-sampling bias or modeling embedding uncertainty.

SoftCLIP [12] is closely related in constructing soft targets from intra-modal similarities. FinAl uses within-batch report-similarity ranks to determine target probabilities, making target construction depend on relative ordering rather than absolute similarity magnitudes. Our controlled ablation supports this choice for severity grading (Table 6). Rank-N-Contrast [13] likewise emphasizes ordered representations, but derives its rankings from regression labels. FinAl instead obtains relational supervision from reports, enabling rank-based image–text alignment without severity annotations during pretraining.

3 Methodology

We propose FinAl for preserving fine-grained clinical information in chest X-ray VLP. Paired images and structured reports are encoded into a shared space (Sec. 3.1); report-side similarities define rank-induced soft targets for global alignment (Secs. 3.2–3.3), MSD adapts its temperature (Sec. 3.4), and bidirectional cross-attention provides local alignment (Sec. 3.5). The objectives are combined in Sec. 3.6.

3.1 Feature Extraction

Given chest X-ray images $X^{(v)}$ and structured radiology reports $X^{(l)}$, we encode both modalities into a shared representation space. For the visual stream, we use a ViT-B/16 [14] encoder initialized from M3AE [15] and follow the CARZero [3] preprocessing pipeline. For the language stream, radiology reports are first converted into Structured Radiology Reports (SRR) using SRRG [16]

and then encoded with SRRG-BERT [16]. Detailed preprocessing and encoder configurations are provided in Appendix A.

3.2 Fine Supervision Extraction

Given a batch of radiology reports $X^{(l)}$, report embeddings $E^{(l)}$ are obtained by applying the text encoder followed by a projection layer and l_2 normalization.

$$\tilde{E}^{(l)} = g_\theta^{(l)}(f_\theta^{(l)}(X^{(l)})) \in \mathbb{R}^{B\times D}, \quad E^{(l)} = \frac{\tilde{E}^{(l)}}{\|\tilde{E}^{(l)}\|_2}, \tag{1}$$

where $f_\theta^{(l)}$ is a text encoder, and $g_\theta^{(l)}$ is a projection head. B and D are the batch size and the dimensionality of report embeddings, respectively. Then, an intra-modal semantic similarity matrix is computed via inner products of report embeddings,

$$S^{(l)} = E^{(l)}{E^{(l)}}^T \in \mathbb{R}^{B\times B}, \tag{2}$$

which captures relative semantic proximity between clinical reports within the batch. We use report similarities to construct the targets, while visual similarities are used in MSD to assess cross-modal rank consistency.

To construct these targets, we use the relative ordering of report similarities rather than their absolute magnitudes. For the i-th row of semantic similarity matrix (corresponded to i-th report), we obtain a permutation π_i by sorting the similarity scores in descending order and permutation-based distance matrix $P \in \{0, ..., B-1\}^{B\times B}$ as:

$$\pi_i^{(l)} = argsort(S_{i,:}^{(l)}), \quad P_{i,\pi_i(j)}^{(l)} = j-1, j = 1, .., B. \tag{3}$$

After sorting candidates by report similarity, we scatter their rank distances back to the original batch indices. This ensures that each target entry $\hat{Y}_{ij}$ corresponds to candidate j in the original batch, rather than the candidate at the j-th position in the sorted list, and is aligned with the corresponding cross-modal prediction. The learning signal therefore comes from converting report-similarity ranks into soft targets, rather than from reordering the batch itself. The permutation-based distances are then transformed into soft target distributions $\hat{Y}$ using a row-wise softmin:

$$\hat{Y}_{ij} = softmin(P_{ij}^{(l)}) = \frac{exp(-P_{ij}^{(l)})}{\sum_{k=1}^{B} exp(-P_{ik}^{(l)})} \tag{4}$$

Here, lower values of $P_{ij}^{(l)}$ indicate that report j is semantically closer to report i within the batch. Since each report is most similar to itself after normalization, the paired sample receives the smallest rank distance and therefore the largest target probability after the softmin. The softmin operation converts rank distances into a normalized probability distribution, allowing the rank-based supervision to be used in a cross-entropy-style contrastive objective while maintaining numerical stability.

3.3 Fine-Detail Preserving Global Alignment

During contrastive training, the soft target distribution $\hat{Y}$ is used as supervision for image-text alignment instead of hard one-hot labels. The cross-modal similarity matrix is defined as

$$S^{(m)} = E^{(v)}{E^{(l)}}^T \in \mathbb{R}^{B\times B} \tag{5}$$

where $E^{(v)}$ and $E^{(l)} \in \mathbb{R}^{B\times D}$ are image and report embeddings respectively. Given a temperature $\tau > 0$, the image-to-text and text-to-image prediction distributions are obtained by normalizing the temperature-scaled similarities over candidate reports and candidate images, respectively:

$$q_{ij}^{v\to l} = \frac{\exp\Big(S_{ij}^{(m)}/\tau\Big)}{\sum_{k=1}^{B}\exp\Big(S_{ik}^{(m)}/\tau\Big)}, \qquad q_{ij}^{l\to v} = \frac{\exp\Big(S_{ji}^{(m)}/\tau\Big)}{\sum_{k=1}^{B}\exp\Big(S_{ki}^{(m)}/\tau\Big)}. \tag{6}$$

Here, i denotes the query index and j the candidate index in each direction.

The smoothed global alignment loss is then defined as

$$\mathcal{L}_{\text{smoothed global}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\hat{Y}_{ij}\left(\log q_{ij}^{v\to l} + \log q_{ij}^{l\to v}\right). \tag{7}$$

This objective replaces the conventional hard one-hot contrastive target with the rank-induced soft target $\hat{Y}$. As a result, clinically related non-paired samples receive nonzero target mass and are not uniformly treated as negatives. With the target distribution and temperature held fixed, each directional cross-entropy term contributes a similarity gradient proportional to the difference between the predicted and target probabilities. This contribution favors increasing similarity when the predicted probability is below its target and decreasing similarity when it is above, including for non-paired candidates. We provide a KL-divergence and gradient interpretation of this objective in Appendix A.4.

3.4 Temperature Scheduling with Modality-wise Similarity Discrepancy

Following a recent finding [17] that the role of the temperature varies across training stages, we incorporate an adaptive temperature strategy in our contrastive objective. [17] showed that temperature balances group-wise discrimination for coarse organization with instance-level discrimination for fine-grained differences.

Motivated by this observation, we adjust the temperature based on the consistency of intra-modal similarity structures between vision and language representations. This consistency is measured using Spearman's rank correlation [18] between intra-modal similarity rankings as shown in Eq. 8 and Eq. 9. We use Spearman correlation because our soft targets are constructed from the relative ordering of report similarities rather than their absolute magnitudes. By comparing ranks, this measure is invariant to strictly increasing transformations of the similarities and reduces dependence on cross-modal scale calibration. This choice provides consistency with our rank-based supervision; we present it as a design rationale rather than evidence of optimality over alternative consistency measures. To compute the modality-wise discrepancy, we construct a visual rank-distance matrix $P^{(v)}$ using the same row-wise ranking procedure used for $P^{(l)}$. Specifically, we first compute the visual intra-modal similarity matrix $S^{(v)} = E^{(v)}E^{(v)T}$, sort each row of $S^{(v)}$ in descending order, and assign rank distances as in Eq. (3).

$$\bar{P}_i^{(v)} = \frac{1}{B-1}\sum_{j\neq i} P_{ij}^{(v)}, \quad \bar{P}_i^{(l)} = \frac{1}{B-1}\sum_{j\neq i} P_{ij}^{(l)}, \tag{8}$$

$$\rho_i = \frac{\sum_{j\neq i}(P_{ij}^{(v)} - \bar{P}_i^{(v)})(P_{ij}^{(l)} - \bar{P}_i^{(l)})}{\sqrt{\sum_{j\neq i}(P_{ij}^{(v)} - \bar{P}_i^{(v)})^2}\sqrt{\sum_{j\neq i}(P_{ij}^{(l)} - \bar{P}_i^{(l)})^2}} \tag{9}$$

The coefficient ρ_i is converted into a rank-based discrepancy. Then the average rank-based discrepancy d and the adaptive temperature are computed as

$$d = \frac{1}{B}\sum_{i=1}^{B}(1-\rho_i), \qquad \tau_{\text{MSD}} = \tau_0\left(1 + \frac{d}{d_{\max}}\right). \tag{10}$$

where τ_0 is the base temperature. Since $\rho_i \in [-1, 1]$, we have $d \in [0, 2]$ and set $d_{\max} = 2$. As a result, larger modality discrepancies yield higher temperatures, producing smoother similarity distributions, while smaller discrepancies allow sharper alignment. During training, we set the temperature τ in Eq. 6 to τ_{MSD}, allowing the contrastive sharpness to adapt to the current alignment state between modalities.

3.5 Local Alignment

To further encourage fine-grained cross-modal correspondence, we perform bidirectional cross-attention alignment [4] between normalized image patch embeddings and text token embeddings, as illustrated in Fig. 2. Let $p_b^r \in \mathbb{R}^D$ denote the r-th normalized image patch embedding of the b-th image, where $b = 1, \ldots, B$ and $r = 1, \ldots, R$. Let $t_b^n \in \mathbb{R}^D$ denote the n-th normalized text token

embedding of the paired report, where $n = 1, \ldots, N$. Here, B is the batch size, R is the number of image patches, and N is the number of text tokens. For each image patch p_b^r, we construct its cross-modal text counterpart c_b^r by attending over all text token embeddings:

$$c_b^r = \sum_{n=1}^{N} \alpha_b^{r,n} t_b^n, \qquad \alpha_b^{r,n} = \frac{\exp\left((p_b^r)^\top t_b^n / \tau_{\text{local}}\right)}{\sum_{n'=1}^{N} \exp\left((p_b^r)^\top t_b^{n'} / \tau_{\text{local}}\right)} \tag{11}$$

The patch-to-text local objective pulls each patch embedding p_b^r closer to its cross-modal text counterpart c_b^r, while contrasting it against the counterparts of other patches from the same image-report pair:

$$\mathcal{L}_{p2t} = -\frac{1}{2BR} \sum_{b=1}^{B} \sum_{r=1}^{R} \omega_b^r \left[\log \frac{\exp\left(\text{sim}(p_b^r, c_b^r)/\tau_2\right)}{\sum_{r'=1}^{R} \exp\left(\text{sim}(p_b^r, c_b^{r'})/\tau_2\right)} + \log \frac{\exp\left(\text{sim}(c_b^r, p_b^r)/\tau_2\right)}{\sum_{r'=1}^{R} \exp\left(\text{sim}(c_b^r, p_b^{r'})/\tau_2\right)} \right] \tag{12}$$

The text-to-patch direction is defined analogously. For each text token t_b^n, we construct a cross-modal image counterpart o_b^n by attending over image patch embeddings, and define $\mathcal{L}_{t2p}$ by contrasting t_b^n with o_b^n over the N token positions. Here, $\text{sim}(\cdot, \cdot)$ denotes cosine similarity, τ_{local} is the cross-attention temperature, τ_2 is the local contrastive temperature, and ω_b^r denotes the patch-level relevance weight, defined as the last-layer attention weight from the r-th visual token to the `[CLS]` token, averaged across attention heads. The final local objective is

$$\mathcal{L}_{\text{local}} = \mathcal{L}_{p2t} + \mathcal{L}_{t2p}. \tag{13}$$

3.6 Overall Objective

The final objective combines global and local alignment losses with hyper-parameters λ_1 and λ_2:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{smoothed global}} + \lambda_2 \mathcal{L}_{\text{local}}. \tag{14}$$

4 Experiment

4.1 Benchmarks and Evaluation Protocols

We evaluate fine-grained severity grading and retrieval on EdemaSeverity [19] and three four-level CAD-Chest [20] severity subsets: atelectasis, pulmonary edema, and cardiomegaly. These benchmarks test whether learned representations distinguish ordered clinical states rather than only disease presence. We also evaluate category-level retrieval on CheXpert8x200 [1] to verify that fine-grained alignment does not degrade standard coarse-grained retrieval utility. Dataset construction, label definitions, prompt templates, and auxiliary fine-grained benchmarks are provided in Appendix B and Appendix D.

4.2 Experimental Setup

FinAl is pretrained on MIMIC-CXR image–report pairs using an M3AE-initialized ViT-B/16 visual encoder and SRRG-BERT with SRRG-based report preprocessing. We compare released checkpoints of GLoRIA [5], MGCA [4], SAT [6], MedKLIP [2], CARZero [3], RADZero [9], ASG [7], and MAVL [8], following their model-specific preprocessing pipelines. For FinAl, we select the checkpoint with the lowest validation loss. Unless otherwise specified, pretrained encoders remain frozen.

Zero-shot classification uses class-prompt similarities, retrieval covers the I2I, I2T, and T2I directions, and linear probing trains only a newly initialized classifier on frozen visual features. We report accuracy (Acc), macro-averaged F1 (macro-F1), mean absolute error (MAE), and quadratic weighted kappa (QWK) for zero-shot severity grading; Acc, sensitivity (Sens), precision (Prec), and macro-F1 for linear probing; and recall at rank 1 (Recall@1) for retrieval. Detailed settings are provided in Appendices A–D.

4.3 Training-free Evaluation

We first evaluate training-free use of the pretrained representations through prompt-based severity grading and severity-aware retrieval.

4.3.1 Severity Grading

Table 1 reports zero-shot severity grading results on EdemaSeverity and three CAD-Chest subsets. We evaluate exact severity classification using accuracy and macro-F1, and ordinal prediction quality using MAE and QWK. FinAl achieves the lowest MAE and highest QWK across all four benchmarks, and the highest accuracy and macro-F1 on EdemaSeverity, atelectasis, and pulmonary edema. On cardiomegaly, SAT achieves higher accuracy and macro-F1 than FinAl (44.45% vs. 42.84% and 41.97% vs. 41.72%, respectively), whereas FinAl yields lower MAE (0.749 vs. 0.843) and higher QWK (0.531 vs. 0.422). These results support improved ordinal severity prediction, while showing that gains in exact classification vary by disease. Class-wise F1 scores and confusion matrices for EdemaSeverity are provided in Appendix D.1.

Table 1: Zero-shot severity grading results. Acc and F1 (macro-F1) are reported as percentages. Higher is better for Acc, F1, and QWK; lower is better for MAE. Bold and underlined values indicate the best and second-best results for each dataset and metric.

Dataset	Edema Severity				CAD-Chest Atelectasis				CAD-Chest Pulmonary Edema				CAD-Chest Cardiomegaly			
Model	Acc	F1	MAE	QWK	Acc	F1	MAE	QWK	Acc	F1	MAE	QWK	Acc	F1	MAE	QWK
GLoRIA	12.18	9.96	1.593	0.104	31.53	13.89	1.223	0.079	36.38	27.87	0.854	0.380	31.45	19.93	1.210	0.166
MGCA	31.25	30.01	0.989	0.108	36.28	24.93	1.040	0.281	35.28	31.13	0.966	0.316	38.13	31.97	0.906	0.416
SAT	37.32	34.41	1.035	0.207	35.73	31.65	1.054	0.193	46.22	45.58	0.791	0.437	**44.45**	**41.97**	0.843	0.422
MedKLIP	30.78	18.69	1.249	0.020	21.13	11.17	1.644	-0.001	23.17	13.15	1.497	0.047	20.94	11.68	1.558	-0.020
CARZero	22.64	14.23	1.028	-0.090	21.90	15.81	1.320	-0.105	24.14	13.11	1.092	-0.023	26.45	20.25	1.244	-0.083
ASG	36.19	30.33	1.079	0.117	38.05	33.57	1.039	0.265	45.84	45.83	0.765	0.404	35.49	33.72	1.073	0.174
MAVL	25.81	14.86	0.963	-0.125	22.12	11.95	1.619	0.002	24.01	17.55	1.528	-0.096	27.55	19.21	1.062	0.060
RADZero	18.18	15.71	1.414	0.066	21.13	11.12	1.074	0.047	25.48	13.12	1.020	0.035	22.34	10.82	1.541	0.012
FinAl (ours)	**44.74**	**42.57**	**0.756**	**0.431**	**40.71**	**37.46**	**0.779**	**0.519**	**52.44**	**52.68**	**0.629**	**0.572**	42.84	41.72	**0.749**	**0.531**

4.3.2 Retrieval-based Representation Evaluation

Retrieval evaluates whether the embedding space preserves severity-aware similarity relationships beyond label prediction. Table 2 reports Recall@1 results on fine-grained and coarse-grained benchmarks. FinAl improves most severity-aware retrieval settings while remaining competitive on CheXpert8x200. Recall@5 results are provided in Appendix D.5.

Table 2: Retrieval results using Recall@1 among VLP models. For fine-grained benchmarks, class-wise average recall is reported. CheXpert8x200 is used as a standard coarse-grained retrieval benchmark. Bold and underlined indicate the best and second-best results.

Dataset	Edema Severity			CAD Atelectasis			CAD Pulmonary Edema			CAD Cardiomegaly			CheXpert	
Model	I2I	I2T	T2I	I2I	I2T	T2I	I2I	I2T	T2I	I2I	I2T	T2I	I2I	T2I
GLoRIA	47.72	26.19	34.48	41.48	39.05	42.59	42.81	23.45	29.98	39.90	34.02	38.28	46.25	45.00
MGCA	42.15	33.73	41.08	40.60	47.23	53.10	47.60	42.81	50.67	43.86	47.17	49.45	43.75	45.00
SAT	39.54	44.92	42.51	41.92	58.41	56.31	48.23	52.73	55.05	42.47	55.77	54.08	42.50	**52.50**
MedKLIP	42.71	19.59	20.04	42.81	23.45	29.98	48.23	27.08	27.42	37.25	29.10	27.11	31.25	20.00
CARZero	51.04	26.90	10.96	42.92	23.12	32.63	49.03	27.71	25.57	42.98	24.25	23.14	35.00	15.00
ASG	52.00	47.29	49.83	39.49	57.08	60.18	50.00	52.94	54.84	45.48	56.43	53.42	**55.00**	**52.50**
MAVL	49.26	19.96	29.64	42.59	21.02	31.08	48.49	26.03	21.19	44.60	25.72	26.30	37.50	7.50
RADZero	40.12	37.10	28.72	29.76	27.10	27.88	37.72	25.44	21.70	33.14	18.81	27.85	21.25	10.00
FinAl(ours)	**54.18**	**52.38**	**51.90**	**45.13**	**71.02**	**69.03**	**50.13**	**63.16**	**58.66**	**45.63**	**66.94**	**59.52**	50.00	47.50

4.4 Linear Probing Evaluation

We next evaluate supervised transfer by freezing each pretrained visual encoder and training only a linear classifier. Table 3 presents linear probing results on EdemaSeverity and CAD-Chest severity grading benchmarks. FinAl achieves the best or highly competitive performance across all evaluated datasets, indicating that the gains are reflected in the frozen visual representation rather than being limited to prompt construction or zero-shot inference. Repeated linear probing results on EdemaSeverity, including comparisons with the non-contrastive baselines M3AE [15] and MRM [21], are provided in Appendix D.7.

Table 3: Linear probing results on severity grading. Bold and underlined indicate the best and second-best results.

Dataset	Edema Severity				CAD-Chest Atelectasis				CAD-Chest Pulmonary Edema				CAD-Chest Cardiomegaly			
Model	Acc	Sens	Prec	F1	Acc	Sens	Prec	F1	Acc	Sens	Prec	F1	Acc	Sens	Prec	F1
GLoRIA	59.38	54.55	57.01	55.24	31.22	31.38	24.21	23.77	36.66	36.66	39.73	34.35	43.08	43.08	42.01	43.09
MGCA	<u>60.92</u>	<u>54.90</u>	58.90	<u>56.17</u>	<u>50.79</u>	<u>50.78</u>	<u>51.14</u>	<u>50.65</u>	<u>58.61</u>	<u>58.61</u>	<u>58.43</u>	<u>57.90</u>	<u>51.69</u>	<u>51.69</u>	<u>50.36</u>	<u>50.87</u>
SAT	57.00	51.30	**61.56**	52.57	32.28	32.44	29.03	25.39	33.74	33.74	28.11	31.65	37.50	37.50	29.54	30.83
MedKLIP	57.84	49.21	55.26	49.86	43.90	39.03	41.77	42.69	39.75	39.75	38.79	38.45	42.12	42.12	40.93	42.01
CARZero	59.24	50.85	58.66	52.63	30.69	30.85	23.89	23.43	38.32	38.32	32.38	36.58	36.82	36.82	38.25	34.07
FinAl(ours)	**62.39**	**55.95**	<u>60.38</u>	**57.35**	**55.56**	**55.47**	**57.93**	**51.46**	**61.48**	**61.48**	**60.64**	**60.75**	**54.05**	**54.05**	**54.10**	**54.04**

4.5 Analysis of Design Choices

4.5.1 Loss Components

Table 4 compares hard one-hot and rank-induced global alignment, each with and without local alignment. In this comparison, adding local alignment decreases macro-F1 from 41.29% to 39.16% with hard one-hot targets, whereas it increases macro-F1 from 40.61% to 42.57% with rank-induced targets. The combination of rank-induced global alignment and local alignment achieves the best results across all four reported metrics.

4.5.2 Temperature Scheduling

Table 5 compares a fixed temperature, a learned temperature, and MSD-based temperature scheduling. The learned-temperature baseline uses a single learnable logit-scale parameter shared by the image-to-text and text-to-image losses, with an initial temperature of 0.07. The learned-temperature variant achieves the highest accuracy, whereas MSD achieves the highest sensitivity, precision, and macro-F1. These results indicate modest, metric-dependent benefits from MSD as a supporting mechanism for adapting alignment sharpness.

Table 4: Loss ablation on EdemaSeverity.

Alignment		Edema Severity			
Fine	Local	Acc	Sens	Prec	F1
		43.78	45.04	42.75	<u>41.29</u>
	✓	40.36	44.26	42.92	39.16
✓		<u>44.72</u>	<u>45.69</u>	<u>43.18</u>	40.61
✓	✓	**44.74**	**46.55**	**43.91**	**42.57**

Table 5: Temperature scheduling ablation on EdemaSeverity (zero-shot classification).

Scheduling	Acc	Sens	Prec	F1
Fixed ($\tau = 0.07$)	44.34	<u>46.32</u>	41.98	40.98
Learned	**45.15**	45.59	<u>42.20</u>	<u>41.49</u>
MSD (Ours)	<u>44.74</u>	**46.55**	**43.91**	**42.57**

4.5.3 Target Construction

To examine the effect of target construction, we compare hard one-hot targets, cosine-based soft targets, and the proposed rank-induced targets, while holding the architecture, pretraining data, report encoder, visual initialization, and training schedule fixed. All variants retain the same local alignment objective, loss weights, and MSD-based temperature scheduling; only the global alignment target construction is changed. For the cosine-based variant, we apply softmax directly to report–report cosine similarities without rank conversion or temperature scaling. MSD-based temperature scaling is applied only to the cross-modal prediction logits. Table 6 reports zero-shot severity grading results on EdemaSeverity. Rank-induced supervision improves both nominal and ordinal metrics over the two alternatives, achieving 44.74% accuracy and 42.57% macro-F1, with an MAE of 0.76 and a QWK of 0.43. These results support using relative report-similarity ordering rather than raw similarity magnitudes in this setting, and show that target construction matters even when the report encoder is shared. We additionally evaluate FinAl with BioClinicalBERT, a widely used text-encoder initialization in medical VLP [5, 4, 6], using original reports and adapted preprocessing (Appendix D.8).

4.6 Representation Analysis

Report-derived supervision. We first examine whether report similarities contain severity-related relational information on EdemaSeverity. Mean report–report cosine similarity is 0.1472, 0.1355,

Table 6: Target-construction ablation on EdemaSeverity zero-shot severity grading. Accuracy and macro-F1 are reported as percentages. Bold and underlined values indicate the best and second-best results.

Target	Acc ↑	Macro-F1 ↑	MAE ↓	QWK ↑
Hard one-hot	<u>40.36</u>	<u>39.16</u>	<u>0.85</u>	<u>0.33</u>
Cosine-based soft target	37.49	27.59	0.98	0.23
Rank-induced (FinAl)	**44.74**	**42.57**	**0.76**	**0.43**

0.0889, and 0.0258 for absolute severity-label differences of 0, 1, 2, and 3, respectively. Similarity shows a modest positive association with severity proximity, defined as the negative absolute label difference (Spearman $\rho = 0.2039$). For non-paired samples, near-severity pairs (label differences 0 or 1) receive higher mean rank-induced target probability than far-severity pairs (differences 2 or 3): 0.001786 versus 0.000772. The difference between these group means (near minus far) is 0.001014. Hard one-hot targets assign zero probability to all non-paired samples. Severity labels are used only for post-hoc analysis, not to construct report similarities or training targets. Details of the pair construction and statistical aggregation are provided in Appendix D.9. A separate analysis of SAT-style threshold-based pair assignments is provided in Appendix D.10.

Transfer of report-induced structure. We compute rank preservation on the full EdemaSeverity evaluation set, using the same studies for all variants without additional subsampling. We use the SRRG-BERT report embeddings to define a common report-induced reference ranking for all controlled variants. For each anchor study, we rank all non-self candidate studies by report–report cosine similarity and compare this fixed reference ranking with the image–image and image–text rankings obtained from each trained variant over the same candidates. Rank preservation is the Spearman correlation between the reference ranking and each representation ranking, averaged across anchors, and does not require severity labels. Table 7 shows that FinAl achieves higher I2I and I2T rank preservation than both controlled alternatives. As a separate ordinal check, FinAl also produces a larger image–text near–far similarity gap and a higher correlation with negative severity distance. Together, these results support improved agreement between report–report and image–image/image–text rankings, alongside a stronger association between image–text similarity and ordinal severity proximity.

Table 7: Quantitative representation analysis on EdemaSeverity. Rank preservation measures agreement with the report-induced candidate ranking, averaged across anchors, without using severity labels. The image–text (I2T) near–far gap is the mean cosine similarity for severity distances 0/1 minus that for distances 2/3; I2T severity correlation is Spearman correlation with negative severity distance. These latter two metrics use severity labels only for post-hoc evaluation. Bold and underlined values indicate the best and second-best results.

Target	Rank preservation ↑		Ordinal analysis ↑	
	I2I	I2T	I2T near–far gap	I2T severity corr.
Hard one-hot	<u>0.5341</u>	<u>0.7239</u>	<u>0.0585</u>	<u>0.1650</u>
Cosine-based soft target	0.4466	0.5401	0.0092	0.1196
Rank-induced (FinAl)	**0.6086**	**0.7863**	**0.0719**	**0.1895**

Qualitative visualization. We complement these quantitative analyses with qualitative visualizations. Figure 3(a) visualizes image and text representations using PCA on EdemaSeverity. Several baseline methods exhibit clear modality gaps or inconsistent overlap between image and text representations of the same severity level. In contrast, FinAl shows more consistent cross-modal overlap across severity levels, suggesting improved alignment of the multimodal representation space.

Figure 3(b) further analyzes severity-wise cross-modal similarity patterns by aggregating image-text cosine similarities over severity class pairs. CARZero exhibits nearly uniform similarity values across report severity levels, indicating limited sensitivity to disease severity. FinAl produces a more structured similarity matrix, with higher similarity concentrated around severity-aligned and adjacent pairs and lower similarity for clinically distant severity pairs. Together, these analyses suggest that FinAl better preserves fine-grained clinical semantics in the learned representation space.

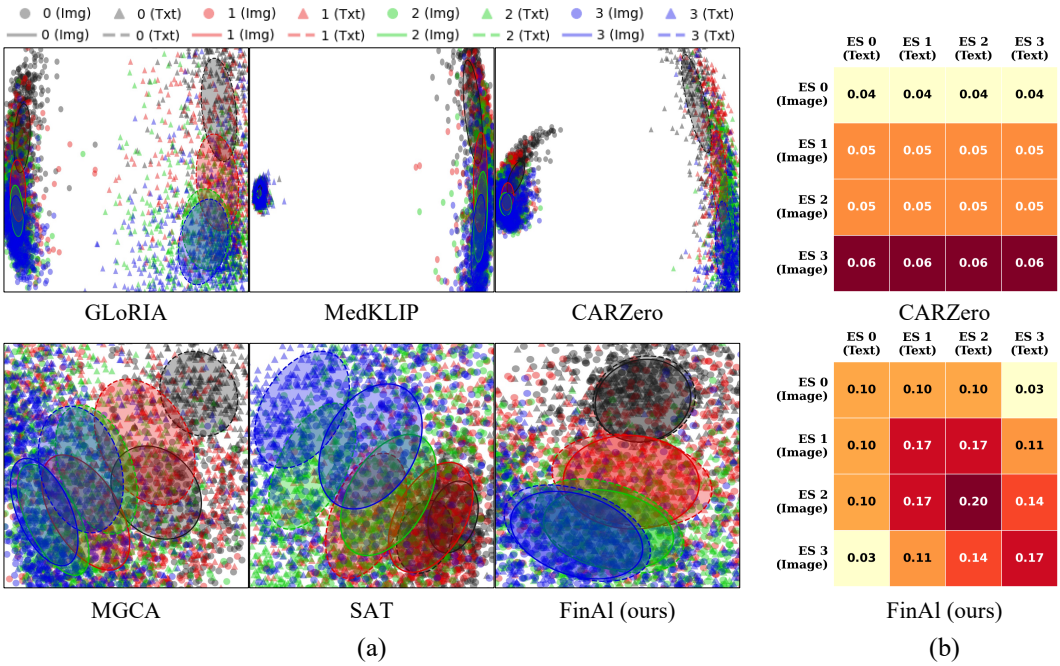


Figure 3: Representation analysis on EdemaSeverity. (a) Multi-modal PCA visualization. Severity-wise distributions are summarized using confidence ellipses; dashed and solid ellipses denote text and image features, respectively. (b) Severity-wise cross-modal similarity matrices between image and text representations.

5 Conclusion

We presented FinAl, a medical vision–language pretraining framework for preserving fine-grained clinical information such as disease severity. FinAl converts within-batch report-similarity rankings into soft targets for global image–text alignment, providing graded supervision for non-paired samples without requiring severity annotations during pretraining. Local alignment and MSD-based adaptation of the cross-modal prediction temperature complement this global objective. On the four evaluated severity benchmarks, FinAl achieves the lowest MAE and highest QWK among the compared models, with gains also observed in severity-aware retrieval and frozen-encoder linear probing. Controlled target-construction ablations on EdemaSeverity further support the benefit of rank-induced supervision over hard one-hot and cosine-based soft targets. FinAl also remains competitive on category-level retrieval. These findings highlight the value of learning from inter-report relationships, beyond matched image–report correspondence, for preserving fine-grained clinical structure in medical VLP representations.

6 Limitations and Future Works

Our experiments primarily assess whether representations learned through medical vision–language pretraining preserve fine-grained clinical information, particularly disease severity. We evaluate this on chest X-ray benchmarks through zero-shot severity grading, severity-aware retrieval, and frozen-encoder linear probing. The effects of report-content and structuring errors on report-derived similarity rankings, rank-induced targets, and downstream performance have not been directly evaluated. Future work could extend the evaluation to other imaging modalities, clinical attributes, and institutions, and examine FinAl as an initialization for task-specific full fine-tuning. FinAl may support more informative medical imaging representations, but report-derived errors could propagate downstream, motivating external validation and clinician oversight before clinical use.

Acknowledgments and Disclosure of Funding

This work was supported by a National Research Foundation of Korea (NRF) grant (RS-2025-00519376) funded by the Korean government (Ministry of Science and ICT, MSIT), the MSIT-funded Bio&Medical Technology Development Program of the NRF (RS-2023-00220408), an MSIT-funded Institute of Information & Communications Technology Planning & Evaluation (IITP) grant for the Artificial Intelligence Graduate School Program at Chung-Ang University (RS-2021-II211341), and an MSIT-funded IITP-ITRC (Information Technology Research Center) grant (IITP-2026-2024-00437102).

References

[1] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022.

[2] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 21372–21383, 2023.

[3] Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11137–11146, 2024.

[4] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in neural information processing systems*, 35:33536–33549, 2022.

[5] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3942–3951, 2021.

[6] Bo Liu, Donghuan Lu, Dong Wei, Xian Wu, Yan Wang, Yu Zhang, and Yefeng Zheng. Improving medical vision-language contrastive pretraining with semantics-aware triage. *IEEE Transactions on Medical Imaging*, 42(12):3579–3589, 2023.

[7] Qingqiu Li, Xiaohan Yan, Jilan Xu, Runtian Yuan, Yuejie Zhang, Rui Feng, Quanli Shen, Xiaobo Zhang, and Shujun Wang. Anatomical structure-guided medical vision-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 80–90. Springer, 2024.

[8] Vu Minh Hieu Phan, Yutong Xie, Yuankai Qi, Lingqiao Liu, Liyang Liu, Bowen Zhang, Zhibin Liao, Qi Wu, Minh-Son To, and Johan W Verjans. Decomposing disease descriptions for enhanced pathology detection: A multi-aspect vision-language pre-training framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11492–11501, 2024.

[9] Jonggwon Park, Byungmu Yoon, Soobum Kim, and Kyoyun Choi. RadZero: Similarity-based cross-attention for explainable vision-language alignment in chest x-ray with zero-shot multi-task capability. *Advances in Neural Information Processing Systems*, 38:56008–56034, 2025.

[10] Ching-Yao Chuang, Joshua W. Robinson, Yen-Chen Lin, Antonio Torralba, and Stefanie Jegelka. Debiased contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 8765–8775, 2020.

[11] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8411–8420. IEEE, 2021.

[12] Yuting Gao, Jinfeng Liu, Zihan Xu, Tong Wu, Enwei Zhang, Ke Li, Jie Yang, Wei Liu, and Xing Sun. SoftCLIP: Softer cross-modal alignment makes CLIP stronger. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1860–1868, 2024.

[13] Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-n-contrast: Learning continuous representations for regression. *Advances in Neural Information Processing Systems*, 36:17882–17903, 2023.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[15] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022.

[16] Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, et al. Automated structured radiology report generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26813–26829, 2025.

[17] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *International Conference on Learning Representations*, 2023.

[18] Jerrold H Zar. Spearman rank correlation. *Encyclopedia of biostatistics*, 7, 2005.

[19] Steven Horng, Ruizhi Liao, Xin Wang, Sandeep Dalal, Polina Golland, and Seth J Berkowitz. Deep learning to quantify pulmonary edema in chest radiographs. *Radiology: Artificial Intelligence*, 3(2):e190228, 2021.

[20] Mengliang Zhang, Xinyue Hu, Lin Gu, Tatsuya Harada, Kazuma Kobayashi, Ronald Summers, and Yingying Zhu. CAD-Chest: Comprehensive Annotation of Diseases based on MIMIC-CXR Radiology Report. *PhysioNet*, December 2023. Version 1.0.

[21] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[22] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[23] Ruizhi Liao, Jonathan Rubin, Grace Lam, Seth Berkowitz, Sandeep Dalal, William Wells, Steven Horng, and Polina Golland. Semi-supervised learning for quantification of pulmonary edema in chest x-ray images. *arXiv preprint arXiv:1902.10785*, 2019.

[24] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[25] Wei Shen Lim, Menno M Van der Eerden, R Laing, Wim G Boersma, Noel Karalus, George I Town, SA Lewis, and J. T. Macfarlane. Defining community acquired pneumonia severity on presentation to hospital: an international derivation and validation study. *Thorax*, 58(5):377–382, 2003.

[26] Carl Marincowitz, Laura Sutton, Tony Stone, Richard Pilbery, Richard Campbell, Benjamin Thomas, Janette Turner, Peter A Bath, Fiona Bell, Katie Biggs, et al. Prognostic accuracy of triage tools for adults with suspected covid-19 in a prehospital setting: an observational cohort study. *Emergency Medicine Journal*, 39(4):317–324, 2022.

[27] Yin-Ting Lin, Ko-Ming Lin, Kai-Hsiang Wu, and Frank Lien. Enhancing pneumonia prognosis in the emergency department: a novel machine learning approach using complete blood count and differential leukocyte count combined with curb-65 score. *BMC Medical Informatics and Decision Making*, 24(1):118, 2024.

[28] Zihao Zhao, Frederik Hauke, Juliana de Castilhos, Sven Nebelung, and Daniel Truhn. Can agents distinguish visually hard-to-separate diseases in a zero-shot setting? a pilot study. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2026*, pages 78–88. Springer Nature Switzerland, 2027.

[29] Joseph Paul Cohen, Paul Bertin, and Vincent Frappier. Chester: A web delivered locally computed chest x-ray disease prediction system. *arXiv preprint arXiv:1901.11210*, 2019.

[30] Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical imaging with deep learning*, pages 74–104. PMLR, 2021.

[31] Fernando Perez-Garcia, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Matthew P Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7(1):119–130, 2025.

A Feature Extraction Detail

A.1 Visual Feature Extraction

For visual feature extraction, we follow the same image preprocessing pipeline and vision encoder configuration as CARZero [3]. Each input image is resized to 256×256 with zero padding to preserve aspect ratio, and converted to a three-channel input. During training, a random crop of size 224×224 is applied. Pixel values are scaled to the [0,1] range and normalized using the statistics adopted in M3AE [15] pretraining, with a mean of 0.4978 and a standard deviation of 0.2449. We employ a ViT-B/16 [14] architecture as the vision encoder, initialized with weights pretrained using M3AE on the MIMIC-CXR-JPG [22] dataset.

A.2 Lingual Feature Extraction

Structured Radiology Reports (SRR) [16], obtained by applying SRRG [16] to MIMIC-CXR radiology reports, are used as the textual input. SRR provides a structured representation of radiology reports that explicitly organizes clinically relevant information.

Preprocessing is performed separately for the Findings and Impression sections, which are known to contain the most critical diagnostic descriptions. Sentences are extracted from each section, where bullet-point items are used for the Findings and numbered items are used for the Impression. Non-ASCII characters are removed, and only sentences with a minimum length of three characters are retained. The processed sentences from both sections are then merged, and duplicate sentences are removed.

The resulting SRR inputs are encoded using a pretrained SRRG-BERT [16] model, which follows a BERT-based architecture and produces contextualized report embeddings.

A.3 Implementation Details

FinAl is implemented in PyTorch. Pretraining is performed using the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.05. The batch size is set to 256, and FinAl is pretrained for 50 epochs. A cosine annealing learning rate scheduler with warmup restarts is applied during pretraining. The vision encoder is initialized from M3AE [15, 3] and receives image inputs of size 224×224 after preprocessing. The text encoder is initialized from SRRG-BERT-Leaves [16] and processes tokenized text with a maximum sequence length of 128. Both encoders are trained jointly during pre-training. We employ a two-layer MLP projection head with a hidden dimension of 1024, following the projection architecture used in MGCA [4]. Image normalization follows the preprocessing scheme described in M3AE. We use loss weights $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$ for global and local alignment, respectively. The MSD base temperature is $\tau_0 = 0.07$. Following the official implementation of MGCA [4], we set the local cross-attention temperature τ_{local} and the local contrastive temperature τ_2 to 0.1.

Compute resources. All FinAl pretraining experiments were conducted on an internal server with two NVIDIA A100 GPUs. A single pretraining run took approximately 80 hours. Training-free evaluations, including zero-shot classification and retrieval, did not involve additional model optimization and required a couple of hours in total. Linear probing experiments were conducted on the same hardware; each probing run required less than an hour, depending on the benchmark. This estimate does not include preliminary exploratory runs that were not used in the final reported results.

A.4 Interpretation of the Smoothed Global Alignment Loss

The proposed objective can be interpreted as replacing the conventional hard target

$$Y_{ij}^{\text{hard}} = \mathbf{1}[i = j] \tag{15}$$

Table 8: Pretraining data and encoder configurations of the compared medical VLP models. The data column lists the image–report pretraining dataset; additional text-encoder training data are described below.

Model	VLP pretraining data	Visual encoder	Text encoder
GLoRIA [5]	CheXpert	ResNet-50	BioClinicalBERT
MGCA [4]	MIMIC-CXR-JPG	ViT-B/16	BioClinicalBERT
SAT [6]	MIMIC-CXR	ResNet-50	BioClinicalBERT
MedKLIP [2]	MIMIC-CXR v2	ResNet-50	ClinicalBERT for entity / knowledge-description encoding
CARZero [3]	MIMIC-CXR	ViT-B/16 (M3AE initialization)	BioClinicalMPBERT
ASG [7]	MIMIC-CXR	ViT-B/16	BioClinicalBERT
MAVL [8]	MIMIC-CXR v2	ResNet-50	BioClinicalBERT
RADZero [9]	MIMIC-CXR	XrayDINOv2	MPNet (all-mpnet-base-v2)
FinAl (Ours)	MIMIC-CXR-JPG	ViT-B/16 (M3AE initialization)	SRRG-BERT-Leaves initialization

with the rank-induced soft target $\hat{Y}_{ij}$. For clarity, consider the image-to-text direction and write $q_{ij} = q_{ij}^{v\to l}$. For each query i, the cross-entropy term satisfies

$$\mathcal{L}_i = -\sum_{j=1}^{B} \hat{Y}_{ij} \log q_{ij} = H(\hat{Y}_{i,:}) + D_{\mathrm{KL}}\left(\hat{Y}_{i,:} \,\|\, q_{i,:}\right). \tag{16}$$

Thus, minimizing this term encourages the cross-modal prediction distribution to match the report-induced semantic neighborhood distribution rather than only the paired positive sample. Holding the target distribution and temperature fixed, the partial derivative with respect to the unscaled cross-modal similarity is

$$\frac{\partial \mathcal{L}_i}{\partial S_{ij}^{(m)}} = \frac{q_{ij} - \hat{Y}_{ij}}{\tau}. \tag{17}$$

The text-to-image term has the analogous form after exchanging the image and report indices. For non-paired samples, the gradient depends on the difference between the predicted probability and the rank-induced target, rather than on the predicted probability alone.

A.5 Pretrained Models and Comparison Settings

Table 8 summarizes the pretraining data and encoder configurations of the medical VLP models used in our main comparisons. For the baselines, we use checkpoints released by the original authors and follow their model-specific preprocessing pipelines. For FinAl, we select the checkpoint with the lowest validation loss. Within each downstream comparison, all models are evaluated using the same benchmarks, splits, labels, and metrics, while retaining their respective pretrained configurations.

CARZero uses BioClinicalMPBERT fine-tuned on MIMIC-CXR and PadChest reports. For FinAl, image preprocessing follows CARZero and M3AE, and the textual input consists of SRRG-structured reports. These released-checkpoint comparisons evaluate complete pretrained systems rather than isolating the training objective. The effect of target construction is examined separately under matched training conditions in Table 6.

Asset licenses and access. All datasets and pretrained checkpoints used in this work are publicly or officially accessible from their original sources and were used in accordance with their respective licenses and terms of use. MIMIC-CXR-JPG and CAD-Chest were accessed under the PhysioNet Credentialed Health Data License and the associated Data Use Agreement. We do not redistribute restricted raw data or third-party checkpoints.

Severity	Keyword
0	“no edema"
	“no vascular congestion"
1	“vascular congestion"
	“hilar engorgement"
2	“interstitial edema"
	“Kerley B"
3	“Alveolar edema"
	“parenchymal opacity"

Table 9: The examples of keywords for each edema severity used when collecting EdemaSeverity benchmark.

B Evaluation Dataset

We use EdemaSeverity [19] and three four-level severity subsets constructed from CAD-Chest [20] as fine-grained benchmarks.

EdemaSeverity [19] collected edema samples for each severity from MIMIC-CXR [22]. To sample each severity, they used a keyword-matching natural language processing method based on regular expressions [23] applied to clinical reports. The representative keywords for each severity are listed in Tab. 9.

Some samples labeled as severity 3 exhibited parenchymal opacity; however, the corresponding reports explicitly stated that pulmonary edema was absent or resolved, with the opacity attributed to consolidation. These cases were initially labeled as severe due to regex-based matching [23] of the term “parenchymal opacity”, and a total of 65 samples were subsequently relabeled as severity 0.

CAD-Chest [20] is a large-scale benchmark covering a wide range of thoracic diseases. From this dataset, we focus on three clinically important pathologies with sufficient sample sizes, namely atelectasis, pulmonary edema and cardiomegaly. For each pathology, we map the No, Mild, Moderate, and Severe groups in CAD-Chest to severity labels 0, 1, 2, and 3, respectively. Since pulmonary edema and cardiomegaly are constructed using stratified sampling with equal class proportions, accuracy and sensitivity (recall) yield identical values in Tab. 3.

CheXpert8x200 [1] is a chest X-ray retrieval benchmark which consists of eight chest abnormality categories, including atelectasis, cardiomegaly, edema, fracture, pleural effusion, pneumonia, pneumothorax, and no finding. For each category, 10 expert-annotated query images and 200 candidate images are included. All images are from CheXpert [24] and especially selected to be positive for only a single category to reduce inter-category ambiguity. In contrast to the fine-grained retrieval setting described earlier, this dataset is used to evaluate retrieval performance at a coarse-grained, category-level setting.

Pneumonia Mortality Risk. The benchmark labels for the Pneumonia mortality risk prediction task were generated using a CURB-65-based rule. CURB-65 is a clinical severity score for community-acquired pneumonia based on five binary criteria: altered mental status (AMS), blood urea nitrogen (BUN), respiratory rate (RR), blood pressure (BP), and age [25]. In our dataset, AMS is operationalized using a Glasgow Coma Scale (GCS)-based criterion, following prior works [26, 27] that use GCS as a proxy for impaired consciousness. For BP, we use systolic blood pressure (SBP) and diastolic blood pressure (DBP). Each satisfied criterion contributes one point. The five criteria and their operational definitions are summarized in Table 10.

Let S_i denote the CURB-65-based score for case i. For pneumonia-positive cases, the score was computed as:

$$\begin{aligned} S_i = \mathbf{1}\{\mathrm{AMS}_i\} + \mathbf{1}\{\mathrm{BUN}_i > 19\} + \mathbf{1}\{\mathrm{RR}_i \geq 30\} \\ + \mathbf{1}\{\mathrm{SBP}_i < 90 \vee \mathrm{DBP}_i \leq 60\} + \mathbf{1}\{\mathrm{Age}_i \geq 65\}, \end{aligned} \tag{18}$$

Table 10: CURB-65-based criteria used for pneumonia risk label construction. Each satisfied criterion contributes one point.

Component	Criterion
AMS	GCS-based altered mental status
BUN	> 19 mg/dL
RR	Respiratory rate ≥ 30/min
BP	SBP < 90 mmHg or DBP ≤ 60 mmHg
Age	≥ 65 years

where $\mathbf{1}\{\cdot\}$ is an indicator function that returns 1 if the condition is satisfied and 0 otherwise. The resulting score ranges from 0 to 5. The final benchmark label y_i was assigned as follows:

$$y_i = \begin{cases} 0, & \text{if pneumonia is absent,} \\ 1, & \text{if pneumonia is present and } S_i \in \{0, 1, 2\}, \\ 2, & \text{if pneumonia is present and } S_i \geq 3. \end{cases}$$

Thus, label 0 corresponds to no risk, label 1 corresponds to low-risk pneumonia, and label 2 corresponds to high-risk pneumonia. These labels were used as rule-based reference labels for model evaluation and should be interpreted as risk strata derived from clinical criteria rather than adjudicated mortality outcomes.

Edema vs Pneumonia [28] was originally curated from the MIMIC-CXR chest radiograph dataset to evaluate whether multimodal agents can distinguish two visually confounded thoracic conditions using images alone. Pulmonary edema and pneumonia can both present with overlapping radiographic findings, such as lung opacities and diffuse haziness, although they differ substantially in etiology and clinical management. In the original dataset construction, the authors used report-derived labels and applied an exclusive-or criterion to retain studies labeled as either edema or pneumonia, but not both. This produced 2,907 candidate studies. To reduce label noise from uncertain radiology reports, they further excluded studies containing low-confidence diagnostic expressions, such as "can not be excluded." The final evaluation set consisted of 1,739 studies, including 878 edema cases and 861 pneumonia cases. We used this dataset as-is for evaluation. For the zero-shot evaluation, we follow the original image-only binary classification setting without task-specific fine-tuning, external tools, annotated supervision, or clinical metadata. As noted in the original study, this dataset should be interpreted as a controlled binary proxy benchmark rather than a fully validated clinical diagnostic dataset. The labels are derived from radiology reports and therefore reflect report impressions rather than an independent clinical reference standard, such as CT confirmation or clinical adjudication. In addition, the exclusive-or formulation does not reflect the possibility that edema and pneumonia may co-occur in real clinical practice.

B.1 Prompt Templates

For severity grading, we use the template "There is [severity] [finding].", where [severity] is selected from no, mild, moderate, and severe. The finding term is set according to the target benchmark, e.g., edema, atelectasis, pulmonary edema, or cardiomegaly. For example, pulmonary edema severity grading uses prompts such as "There is no edema.", "There is mild edema.", "There is moderate edema.", and "There is severe edema." The same prompt templates are used for all compared VLP models.

For the additional prompt-based zero-shot classification tasks, we evaluate both short and long prompt variants. Short prompts use concise class descriptions, whereas long prompts provide more explicit clinical context for the corresponding class. Table 11 summarizes the prompt templates used for pneumonia mortality risk prediction, and Table 12 summarizes the prompt templates used for edema-vs-pneumonia differential diagnosis. The same prompt templates are used for all compared VLP models.

Table 11: Prompt templates for pneumonia mortality risk prediction.

Class	Short prompt	Long prompt
No pneumonia / no risk	`there is no pneumonia`	`no radiographic evidence of pneumonia.`
Low-risk pneumonia	`there is mild pneumonia`	`findings are consistent with pneumonia with low clinical severity.`
High-risk pneumonia	`there is severe pneumonia`	`findings are consistent with pneumonia with high clinical severity.`

Table 12: Prompt templates for edema-vs-pneumonia differential diagnosis.

Class	Short prompt	Long prompt
Pneumonia	`there is pneumonia.`	`this chest X-ray shows pneumonia.`
Pulmonary edema	`there is pulmonary edema.`	`this chest X-ray shows pulmonary edema.`

B.2 Evaluation Protocol and Metrics

For prompt-based zero-shot classification, each class label is converted into a text prompt and encoded by the text encoder. Given an input image, we compute cosine similarity between the image embedding and each class prompt embedding, and assign the class with the highest similarity.

For retrieval evaluation, we consider image-to-image (I2I), image-to-text (I2T), and text-to-image (T2I) settings. In I2T retrieval, each image is used as a query and candidate reports are ranked by cosine similarity. T2I retrieval is computed analogously using each report as a query. For I2I retrieval, images are ranked according to cosine similarity in the visual embedding space. For severity-aware retrieval, a retrieved sample is considered relevant when it has the same severity label as the query. We compute recall within each severity class and average across classes to reduce the effect of class imbalance.

For the zero-shot severity-grading comparison in Table 1, we report accuracy (Acc), macro-averaged F1-score (F1), mean absolute error (MAE), and quadratic weighted kappa (QWK). MAE and QWK additionally account for the ordered structure of the severity labels. For the single-run linear probing comparison in Table 3, we report Acc, macro-averaged sensitivity (Sens.), macro-averaged precision (Prec.), and macro-F1. Repeated linear probing results on EdemaSeverity are reported as the mean and standard deviation of Acc, macro-F1, MAE, and QWK in Appendix D.7. The main paper reports Recall@1 for retrieval, and Appendix D.5 reports Recall@5.

C Linear Probing Hyperparameter Search Space

For VLM-based linear probing experiments, we freeze the pretrained image encoder and train only a newly initialized linear classification head on top of the visual representation. All VLM baselines and FinAl are evaluated under the same linear probing protocol. We train each linear classifier for a maximum of 200 epochs using the RAdam optimizer. The batch size is set to 48. For the results in Table 3, the random seed is fixed to 42. We use a learning-rate scheduler with 3 warmup epochs, a cycle length of 30 epochs, and a decay factor of $\gamma = 0.7$. We perform a grid search over learning rate and weight decay. The learning rate is selected from $\{5 \times 10^{-5}, 3 \times 10^{-4}, 5 \times 10^{-4}, 3 \times 10^{-3}, 5 \times 10^{-3}\}$, and the weight decay is selected from $\{10^{-3}, 10^{-4}\}$. This results in 10 hyperparameter configurations for each VLM model and downstream task. For each model and downstream task, we select the linear-classifier checkpoint and hyperparameter configuration with the lowest validation loss.

For vision-only baselines, we follow the same linear probing principle: the pretrained visual encoder is frozen, and only a newly initialized linear classification head is trained on top of the extracted visual representation. Unlike the VLM-based probing experiments, vision-only baselines are trained with a shorter schedule and early stopping. Each linear classifier is trained for a maximum of 50 epochs using the RAdam optimizer. The batch size is set to 48, and the random seed is fixed to 42 for all runs. We use a learning-rate scheduler with 3 warmup epochs, a cycle length of 15 epochs, and a decay factor of $\gamma = 0.7$. Weight decay is fixed to 10^{-3}. We perform a grid search over the learning rate, selected from $\{5 \times 10^{-5}, 3 \times 10^{-4}, 5 \times 10^{-4}, 3 \times 10^{-3}, 5 \times 10^{-3}\}$. This results in 5 hyperparameter configurations for each vision-only baseline and downstream task. Early stopping is applied based on validation performance. Training is stopped if the validation criterion does not improve by at least 10^{-4} for 5 consecutive epochs.

D Additional Analyses and Auxiliary Evaluations

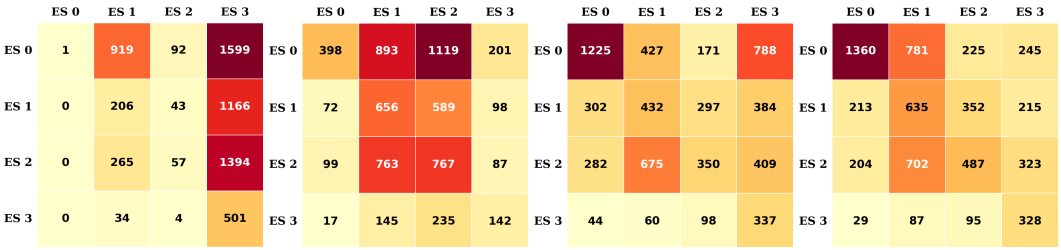
D.1 EdemaSeverity Additional Analysis

Table 13: Additional EdemaSeverity zero-shot analysis with class-wise F1, MAE, and QWK.

Models	ES-0↑	ES-1↑	ES-2↑	ES-3↑	MAE↓	QWK↑
MGCA	24.90	33.88	**34.66**	26.62	0.99	0.11
SAT	54.88	28.71	26.60	27.43	1.03	0.21
FinAl (Ours)	**61.58**	**35.08**	33.88	**39.76**	**0.76**	**0.43**

To complement the aggregate zero-shot severity grading results in Table 1, we examine class-wise F1 scores and confusion matrices on EdemaSeverity. Compared with MGCA and SAT, FinAl achieves higher F1 scores for severity levels 0, 1, and 3, whereas MGCA performs slightly better at severity level 2. This class-wise analysis shows how performance varies across severity levels, complementing the aggregate classification and ordinal metrics reported in the main text.

Figure 4 shows the confusion matrices for zero-shot classification on EdemaSeverity. GLoRIA exhibits a strong tendency to over-predict the highest severity level, indicating poor preservation of graded severity structure. MGCA alleviates this tendency, but many predictions are still concentrated in the middle severity levels, leading to substantial confusion across neighboring classes. SAT improves the recognition of the extreme classes, but confusion among the intermediate severity levels remains considerable. FinAl shows the clearest overall diagonal pattern among the compared VLP baselines. In particular, it achieves higher class-wise F1 for severity levels 0, 1, and 3 than MGCA and SAT, while reducing large errors between clinically distant severity levels compared with GLoRIA and SAT. For FinAl, severity level 2 has the lowest class-wise F1 and is often confused with adjacent levels, especially severity level 1. These error patterns are consistent with the class-wise and ordered-label results in Table 13, where FinAl improves MAE, QWK, and most class-wise F1 scores.



(a) GLoRIA (b) MGCA (c) SAT (d) FinAl (Ours)

Figure 4: Confusion matrices for EdemaSeverity zero-shot classification. Each entry shows the number of samples.

D.2 Fine-grained Differential Diagnosis

To evaluate whether representations distinguish visually and clinically related conditions, we include an edema-vs-pneumonia benchmark inspired by the visually hard-to-separate disease setting in [28]. Unlike severity grading, this task evaluates fine-grained discrimination between two different conditions that can exhibit overlapping radiographic patterns.

Table 14: Edema vs. Pneumonia zero-shot classification.

	short prompt			long prompt		
Models	MAE↓	QWK↑	F1↑	MAE↓	QWK↑	F1↑
MGCA	0.45	0.11	47.34	0.42	0.18	54.10
SAT	0.40	0.21	62.68	0.38	0.25	62.83
FinAl (Ours)	**0.32**	**0.36**	**69.70**	**0.21**	**0.57**	**75.33**

Table 14 reports zero-shot classification results under short and long prompt settings. FinAl outperforms the compared VLP baselines across MAE, QWK, and F1 under both prompt variants,

suggesting that the proposed fine-grained alignment improves discrimination between visually related thoracic conditions.

We also evaluate linear probing on the edema-vs-pneumonia task. As shown in Table 15, FinAl achieves the strongest accuracy and QWK, while matching the best MAE among the compared models. This result supports that FinAl improves supervised transfer to fine-grained differential diagnosis tasks as well as severity grading tasks.

Table 15: Edema vs. Pneumonia linear probing.

Models	ACC↑	MAE↓	QWK↑
MGCA	81.15	0.19	0.62
SAT	<u>83.85</u>	**0.16**	<u>0.68</u>
FinAl (Ours)	**84.23**	**0.16**	**0.69**

D.3 Pneumonia Mortality Risk Prediction

We evaluate pneumonia mortality risk prediction using labels derived from a modified CURB-65 criterion [25]. Non-pneumonia cases are assigned to the no-risk class, while pneumonia cases are grouped into low-risk and high-risk classes based on modified CURB-65 scores. Since this task reflects clinically meaningful risk strata rather than direct observed mortality outcomes, we treat it as a risk-stratified clinical state prediction task.

We evaluate zero-shot pneumonia mortality risk prediction in Table 16. This task extends the evaluation from radiographic severity grading to the risk strata defined in the "Pneumonia Mortality Risk" paragraph in Appendix B. FinAl achieves the lowest MAE and the highest QWK and precision among the compared methods under both short- and long-prompt settings.

Table 16: Zero-shot classification results for pneumonia mortality risk prediction.

	short prompt			long prompt		
Models	MAE↓	QWK↑	Prec.↑	MAE↓	QWK↑	Prec.↑
GLoRIA	0.75	0.13	19.28	0.79	0.06	43.04
MedKLIP	0.77	0.15	28.49	1.15	0.00	28.69
MGCA	<u>0.52</u>	<u>0.49</u>	<u>56.07</u>	0.80	0.06	<u>56.31</u>
SAT	0.63	0.36	54.21	<u>0.55</u>	<u>0.49</u>	53.37
ASG	0.59	0.43	54.94	0.58	0.41	55.74
CARZero	0.94	-0.12	23.61	0.84	0.03	36.65
MAVL	0.73	0.29	30.87	1.05	0.04	21.96
RADZero	0.80	0.21	41.98	0.70	0.31	46.00
FinAl (Ours)	**0.49**	**0.59**	**59.32**	**0.54**	**0.54**	**59.36**

D.4 Comparison with Vision-only Encoders

We additionally compare pretrained visual representations through linear probing and image-to-image retrieval. The linear probing comparison includes the vision-only baselines Dense121 [29], ResNet Auto-encoder [30], and RadDINO [31]. For retrieval, we compare the visual encoders of M3AE [15] and RadDINO with that of FinAl. M3AE is a multimodal pretraining method; only its visual encoder is used in this retrieval evaluation.

Linear probing. Table 17 reports results on the EdemaSeverity linear probing task. FinAl outperforms the vision-only baselines across the reported metrics, including MAE, QWK, accuracy, and F1, suggesting that fine-grained language-derived supervision improves the visual representation for severity-aware clinical prediction.

Retrieval. Table 18 reports image-to-image retrieval results using the visual encoders of M3AE, RadDINO, and FinAl on fine-grained and coarse-grained benchmarks. This comparison evaluates image representations without using text embeddings at retrieval time.

Table 17: Vision-only baseline comparison for EdemaSeverity linear probing task

Models	MIMIC-CXR pre-trained	MAE↓	QWK↑	ACC↑	F1↑
Dense121	✓	0.75	0.39	48.46	37.41
ResNet Auto-encoder	✓	1.00	0.04	42.16	20.19
RadDINO	✓	<u>0.60</u>	<u>0.53</u>	<u>58.19</u>	<u>52.10</u>
FinAl (Ours)	✓	**0.53**	**0.59**	**62.39**	**57.35**

Table 18: Image-to-image retrieval comparison using the visual encoders of M3AE, RadDINO, and FinAl. For fine-grained benchmarks, class-wise average recall is reported.

Model	Edema Severity		CAD Atelectasis		CAD Pulmonary Edema		CAD Cardiomegaly		CheXpert8x200	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
M3AE	42.84	82.85	39.74	84.15	39.74	84.15	35.56	83.03	21.25	68.75
RadDINO	<u>51.20</u>	<u>84.13</u>	**45.21**	**87.13**	<u>45.21</u>	<u>87.13</u>	<u>38.80</u>	<u>84.13</u>	<u>33.75</u>	<u>78.75</u>
FinAl (Ours)	**54.18**	**86.39**	<u>45.13</u>	<u>84.40</u>	**50.13**	**88.31**	**45.63**	**87.95**	**50.00**	**88.75**

Table 19: Retrieval results using Recall@5. For fine-grained benchmarks, class-wise average recall is reported. CheXpert8x200 is used as a standard coarse-grained retrieval benchmark. Bold and underlined indicate the best and second-best results.

Dataset	Edema Severity			CAD Atelectasis			CAD Pulmonary Edema			CAD Cardiomegaly			CheXpert8x200	
Model	I2I	I2T	T2I	I2I	I2T	T2I	I2I	I2T	T2I	I2I	I2T	T2I	I2I	T2I
GLoRIA	85.06	51.07	76.22	85.51	83.08	87.06	84.29	61.28	76.33	84.13	70.17	84.86	85.00	<u>87.50</u>
MGCA	81.26	67.79	80.37	83.30	85.95	91.04	87.72	79.14	90.12	84.86	83.10	88.02	86.25	85.00
SAT	81.87	<u>80.90</u>	82.98	85.07	92.15	93.14	88.35	87.80	89.02	85.60	88.83	<u>93.09</u>	85.00	80.00
MedKLIP	80.93	57.40	60.22	84.29	61.28	76.33	88.10	67.28	75.19	83.84	67.82	64.58	76.25	60.00
CARZero	85.45	71.18	56.93	**86.84**	76.33	77.21	**89.02**	65.73	65.64	<u>87.66</u>	67.60	62.97	83.75	62.50
ASG	<u>86.12</u>	79.60	<u>87.69</u>	<u>86.28</u>	<u>94.03</u>	<u>95.02</u>	88.14	<u>91.04</u>	<u>91.55</u>	86.63	<u>90.23</u>	91.70	**90.00**	**90.00**
MAVL	85.04	67.15	68.19	85.51	69.25	69.25	<u>88.90</u>	84.15	71.32	86.48	78.47	67.38	85.00	65.00
RADZero	84.32	71.82	77.90	79.42	89.38	78.65	85.45	61.61	64.30	80.75	59.44	74.50	68.75	37.50
FinAl(ours)	**86.39**	**86.41**	**89.61**	84.40	**96.35**	**95.91**	88.31	**94.45**	**95.12**	**87.95**	**95.37**	**94.20**	<u>88.75</u>	<u>87.50</u>

D.5 Recall@5 Retrieval Results

Table 19 reports the Recall@5 retrieval results on both fine-grained severity-aware benchmarks and the standard coarse-grained CheXpert8x200 benchmark. The main paper reports Recall@1 in Table 2 for compactness.

D.6 Standard Disease-level Zero-shot Classification

We complement the fine-grained evaluations with disease-level zero-shot classification on RSNA Pneumonia, CheXpert 5×200, ChestX-ray14, and Open-I to assess performance beyond severity-specific tasks. On RSNA, we evaluate binary pneumonia classification using the prompts "there is pneumonia" and "there is no pneumonia", assigning each image to the prompt with the higher image–text similarity. AUROC and AUPRC are computed from the pneumonia score, with pneumonia as the positive class. For CheXpert 5×200, we evaluate five-way single-label classification using the prompt template "there is {disease}". Accuracy and macro-F1 are computed by selecting the class with the highest image–text logit. For AUROC and AUPRC, we apply softmax over the five class logits and macro-average the one-vs-rest evaluations. CARZero [3] is evaluated using its official zero-shot inference procedure, which incorporates local image features in addition to global image–text similarity.

Table 20 shows that FinAl achieves the highest AUROC, AUPRC, accuracy, and macro-F1 among the compared methods on RSNA, although its gains over ASG are modest. On CheXpert 5×200, FinAl achieves the highest accuracy and macro-F1, while CARZero obtains slightly higher macro-AUROC and macro-AUPRC.

On ChestX-ray14 multi-label zero-shot classification, FinAl obtains an AUROC of 0.6490 and an AUPRC of 0.1157, below CARZero (0.7895 and 0.2098 with its official local-feature-based inference)

Table 20: Zero-shot classification results on standard classification tasks in the medical domain. Bold and underlined indicate the best and second-best results.

Dataset	RSNA				CheXpert5x200			
Model	AUROC	AUPRC	Acc	F1	AUROC	AUPRC	Acc	F1
GLoRIA	0.8092	0.6325	0.4133	0.3483	0.6211	0.3010	0.2720	0.1920
SAT	0.7127	0.5356	0.6607	0.6517	0.6026	0.3465	0.3090	0.2783
CARZero	0.8056	0.6213	0.6887	0.6220	**0.8050**	**0.5658**	<u>0.5030</u>	<u>0.4919</u>
ASG	<u>0.8281</u>	<u>0.6909</u>	<u>0.7313</u>	<u>0.7281</u>	0.7613	0.5182	0.4860	0.4784
FinAl(ours)	**0.8331**	**0.6981**	**0.7357**	**0.7311**	<u>0.8040</u>	<u>0.5596</u>	**0.5180**	**0.5000**

Table 21: Repeated linear probing on EdemaSeverity with frozen visual encoders. Results are mean ± standard deviation across four linear probing runs with different random seeds, not repeated VLP pretraining. Acc and F1 (macro-averaged) are percentages. Bold and underlined means indicate the best and second-best results, respectively.

Model	Acc ↑	F1 ↑	MAE ↓	QWK ↑
M3AE	49.84 ± 0.54	45.04 ± 0.73	0.7316 ± 0.0023	0.4069 ± 0.0124
MRM	58.82 ± 0.80	55.67 ± 0.45	0.5605 ± 0.0199	0.5846 ± 0.0120
GLoRIA	59.64 ± 0.19	55.31 ± 0.27	0.5666 ± 0.0033	0.5794 ± 0.0011
MGCA	<u>60.78</u> ± 0.31	<u>55.71</u> ± 0.53	<u>0.5473</u> ± 0.0040	<u>0.5848</u> ± 0.0026
SAT	57.09 ± 0.16	52.44 ± 0.17	0.5613 ± 0.0019	0.5427 ± 0.0017
MedKLIP	57.60 ± 0.25	50.13 ± 0.32	0.6086 ± 0.0030	0.5287 ± 0.0044
CARZero	59.09 ± 0.35	52.79 ± 0.51	0.5912 ± 0.0051	0.5194 ± 0.0080
FinAl (Ours)	**62.080** ± 0.201	**56.805** ± 0.351	**0.5336** ± 0.0047	**0.5911** ± 0.0033

and ASG, but above GLoRIA and SAT in both metrics. On Open-I, CARZero again performs best, while FinAl achieves an accuracy of 0.4823 and an F1-score of 0.1015, exceeding ASG, GLoRIA, and SAT on these two metrics. However, ASG achieves higher AUROC and AUPRC than FinAl. Together with the RSNA and CheXpert results, these evaluations extend the assessment beyond severity-specific tasks while showing that performance varies across datasets and metrics. Our primary claim remains improved fine-grained clinical representation, rather than universal superiority in zero-shot disease classification.

D.7 Repeated Linear Probing

We evaluate repeated linear probing on EdemaSeverity by keeping each pretrained visual encoder frozen and training a newly initialized linear classifier for each run. Table 21 reports the mean and standard deviation of accuracy, macro-F1, MAE, and QWK for eight methods, including the non-contrastive baselines M3AE [15] and MRM [21]. FinAl achieves the highest mean accuracy, macro-F1, and QWK and the lowest mean MAE among the evaluated methods. These results complement the single-run comparisons in Table 3; the reported variability reflects repeated linear-head training rather than repeated VLP pretraining.

D.8 Report Encoder and Preprocessing Sensitivity

We examine whether FinAl is tied to its default report-processing configuration. We choose BioClinicalBERT as an alternative because it is a widely used text-encoder initialization in prior medical VLP models, including GLoRIA [5], MGCA [4], and SAT [6]. The default model uses SRRG-structured reports with SRRG-BERT, whereas the alternative uses original radiology reports without SRRG conversion, together with BioClinicalBERT and preprocessing adapted to that encoder. We therefore compare two report-processing configurations rather than imposing the SRRG-specific input format on both encoders. This comparison changes both the report encoder and preprocessing and does not

Table 22: Report encoder and preprocessing comparison on EdemaSeverity zero-shot severity grading. The BioClinicalBERT variant uses original reports with adapted preprocessing, while the default FinAl uses SRRG-structured reports with SRRG-BERT. Acc, Prec., and Sens. are reported as percentages. Bold and underlined values indicate the best and second-best results.

Model / Report encoder	Acc ↑	Prec. ↑	Sens. ↑	MAE ↓	QWK ↑
SAT	37.32	37.25	40.09	1.03	0.21
FinAl (BioClinicalBERT)	<u>40.17</u>	**44.15**	<u>42.13</u>	<u>0.78</u>	<u>0.37</u>
FinAl (SRRG-BERT)	**44.74**	<u>43.91</u>	**46.55**	**0.76**	**0.43**

isolate the effect of the encoder alone. We evaluate zero-shot severity grading on EdemaSeverity and include SAT as a reference baseline.

Table 22 shows that the BioClinicalBERT variant outperforms SAT on all five reported metrics. The default SRRG-based configuration achieves higher accuracy, sensitivity, and QWK and lower MAE, while the BioClinicalBERT variant achieves higher precision. These results suggest that FinAl can also be used with an alternative report-processing configuration, although performance depends on the combined choice of encoder and preprocessing. This experiment does not establish robustness to report-content noise or SRRG extraction errors.

D.9 Details of Report-derived Supervision Analysis

We use the full EdemaSeverity evaluation set without additional study-level filtering or class balancing. Report embeddings are extracted from the trained FinAl checkpoint used for downstream evaluation, including the text projection head and L2 normalization. The model parameters remain fixed during this analysis. For studies i and j, let t_i and t_j denote their normalized report embeddings and let $y_i, y_j \in \{0, 1, 2, 3\}$ denote their severity labels. We define report similarity and severity distance as

$$s_{ij} = t_i^\top t_j, \qquad d_{ij} = |y_i - y_j|.$$

For report-similarity statistics, we sample 100,000 unordered non-self study pairs from each severity-distance group, yielding 400,000 sampled pairs in total (random seed 42). Let (P) denote the collection of sampled pairs. For each severity distance $d \in \{0, 1, 2, 3\}$, we compute

$$\mathcal{P}_d = \{(i, j) \in \mathcal{P} : d_{ij} = d\}, \qquad \bar{s}_d = \frac{1}{|\mathcal{P}_d|} \sum_{(i,j)\in\mathcal{P}_d} s_{ij}.$$

Each pair contributes equally within its distance group; we do not first average within individual studies or severity classes.

To measure the association between report similarity and severity proximity, we compute Spearman rank correlation between $\{s_{ij}\}_{(i,j)\in\mathcal{P}}$ and $\{-d_{ij}\}_{(i,j)\in\mathcal{P}}$. Tied values are assigned average ranks. The correlation is computed over the pooled pair-level values, not over the four distance-specific means. A positive coefficient indicates that report pairs with closer severity levels tend to have higher similarity.

For target-probability statistics, we construct 500 full batches of 256 studies using repeated random permutations of the evaluation studies with a random seed of 42. Within each batch, every study serves as a reference, and the reports in the same batch form its candidate set. Batch construction does not use severity labels. The batch size matches that used during pretraining.

Within each candidate set, we construct rank distances using the same report-similarity ranking procedure as in training. Let $r_i(j)$ denote the rank distance assigned to candidate j. The target probability is

$$\hat{Y}_{ij} = \frac{\exp(-r_i(j))}{\sum_{k\in\mathcal{C}_i} \exp(-r_i(k))}, \qquad j \in \mathcal{C}_i.$$

The paired entry is included when constructing and normalizing the target distribution. We exclude it only when aggregating non-paired target probabilities and do not renormalize the remaining entries. No temperature scaling is applied to the target distribution.

Table 23: SAT-style triage assignments on EdemaSeverity using BioClinicalBERT report similarities. Values are percentages within each row's pair set. Positive and neutral assignments are reported jointly.

Pair set	Negative (%)	Positive or neutral (%)
All analyzed report pairs	99.84	0.16
Near-severity pairs	99.81	0.19

We pool the non-paired target entries across reference studies and define

$$\mathcal{Q}_{\text{near}} = \{(i,j) : j \in \mathcal{C}_i,\ j \neq i,\ d_{ij} \in \{0,1\}\},$$
$$\mathcal{Q}_{\text{far}} = \{(i,j) : j \in \mathcal{C}_i,\ j \neq i,\ d_{ij} \in \{2,3\}\}.$$

For each group $g \in \{\text{near}, \text{far}\}$, we report the arithmetic mean of its target entries and compute the near–far difference as

$$\bar{p}_g = \frac{1}{|\mathcal{Q}_g|} \sum_{(b,i,j)\in\mathcal{Q}_g} \hat{Y}_{ij}^{(b)}, \qquad \Delta_p = \bar{p}_{\text{near}} - \bar{p}_{\text{far}}.$$

These quantities are mean target probabilities per non-paired entry, rather than the fraction of pairs assigned to a category or the total probability allocated to a group. For hard one-hot targets, all non-paired entries are zero. Severity labels are used only to group pairs and compute post-hoc statistics, not to construct report embeddings, select candidates, or generate targets.

D.10 SAT-style Triage on EdemaSeverity

We examine how SAT-style triage assigns report pairs to positive, neutral, and negative groups on EdemaSeverity. We apply SAT-style fixed thresholds to BioClinicalBERT report similarities, following the triage formulation of SAT [6]. Near-severity pairs have an absolute severity-label difference of 0 or 1. Severity labels are used only to identify this subset for the post-hoc analysis.

Table 23 shows that the positive/neutral assignment rate is slightly higher for near-severity pairs than for all analyzed pairs (0.19% vs. 0.16%). Nevertheless, 99.81% of near-severity pairs are assigned to the negative group under this configuration. FinAl's rank-induced target probabilities are analyzed separately in Section 4.6, where we examine the mean probability assigned to non-paired samples at different severity distances.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state that FinAl aims to improve fine-grained representation fidelity for medical VLP.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Section 6, including dependence on structured radiology reports, possible report-level noise, and the current focus on chest X-ray severity-related benchmarks rather than broader clinical deployment.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theoretical results or theorem-proof statements. The mathematical formulations are used to define the training objective and optimization components.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the model components, training objectives, benchmark construction, evaluation protocols, and hyperparameter search spaces. Additional dataset details, prompt templates, and implementation settings are provided in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The current submission does not include an open public code release. We plan to release the full training, evaluation, and benchmark-construction code upon publication. The newly constructed benchmarks are based on existing public or officially accessible datasets; rather than redistributing restricted raw medical data, we will provide scripts and detailed instructions for reconstructing the evaluation subsets from the official data sources.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the evaluation protocols, baseline usage, preprocessing assumptions, metrics, and hyperparameter search spaces. Additional implementation and dataset construction details are included in the appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 21 reports mean and standard deviation across four linear-probing runs with different random seeds. These statistics reflect variability in linear-head training rather than repeated VLP pretraining. The zero-shot and retrieval evaluations use fixed pretrained checkpoints and benchmark splits.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU type, number of GPUs, approximate wall-clock training time, and compute requirements for pretraining, training-free evaluation, and linear probing in the implementation appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses de-identified public medical imaging datasets and is intended for representation learning research rather than direct clinical deployment. We do not introduce procedures that conflict with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential benefits for fine-grained medical representation learning and acknowledges risks related to report-derived supervision, limited validation scope, and non-readiness for immediate clinical deployment.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a high-risk generative model, scraped dataset, or system intended for direct deployment. The proposed method is evaluated as a research model for medical representation learning.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper credits the datasets, pretrained models, and baseline methods used in the experiments. The appendix summarizes the applicable licenses and access terms, including the PhysioNet license and Data Use Agreement for restricted medical datasets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The submission does not release a new standalone dataset, model, or code asset. All evaluation benchmarks are derived from existing public or officially accessible medical datasets, and the paper documents the construction rules, label definitions, and prompt templates in the appendix. We will release the full training and evaluation code, together with benchmark reconstruction scripts and instructions, upon publication. The reconstruction scripts will require users to obtain the original datasets from their official sources and comply with the corresponding licenses and data-use agreements; we will not redistribute restricted raw medical data.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The work does not involve crowdsourcing experiments or newly collected human-subject data.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work uses public de-identified medical imaging datasets and does not involve new human-subject data collection or intervention.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The proposed method, data construction, training, and evaluation pipeline do not use LLMs. LLM usage in compared baseline methods follows their original implementations and is not introduced by this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.